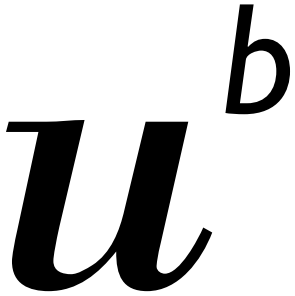


b

UNIVERSITÄT
BERN

UNIVERSITÄT BERN

Research Data Management (RDM)

Data quality and metadata standards

Dr. Olga Churakova, Dr. Gero Schreier

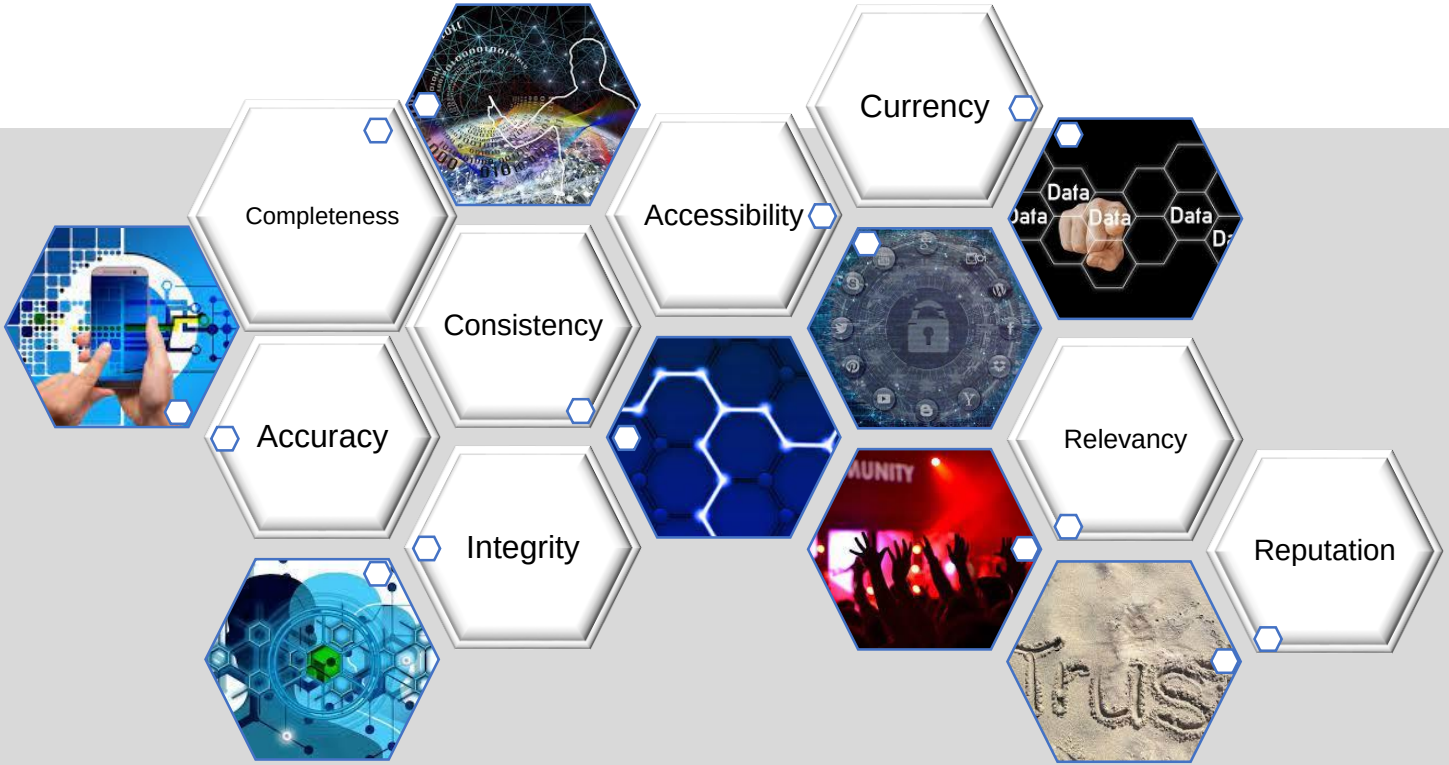
Open Science Team, UB Bern, openscience@ub.unibe.ch

Data quality and metadata standards

Outline

- Data quality dimensions
- Data limitations and data protection
- Towards machine-readable standards
- Metadata standards and controlled vocabularies
- Documentation
- Data quality control & data cleaning
- Good Laboratory Practice Guidelines (GLP)
- Good Clinical Practice (GCP)
- Support

Data quality dimensions



Protection of sensitive data

UNIVERSITÄT BERN

Data protection and sensitive data

Act on Data Protection of the Canton of Bern (KDSG), Art. 2.

For informational purposes in English: Federal Act on Data Protection, Art. 3a, c, d.

Other countries, e.g., Europe (GDPR)

IT-Security awareness @ UniBE

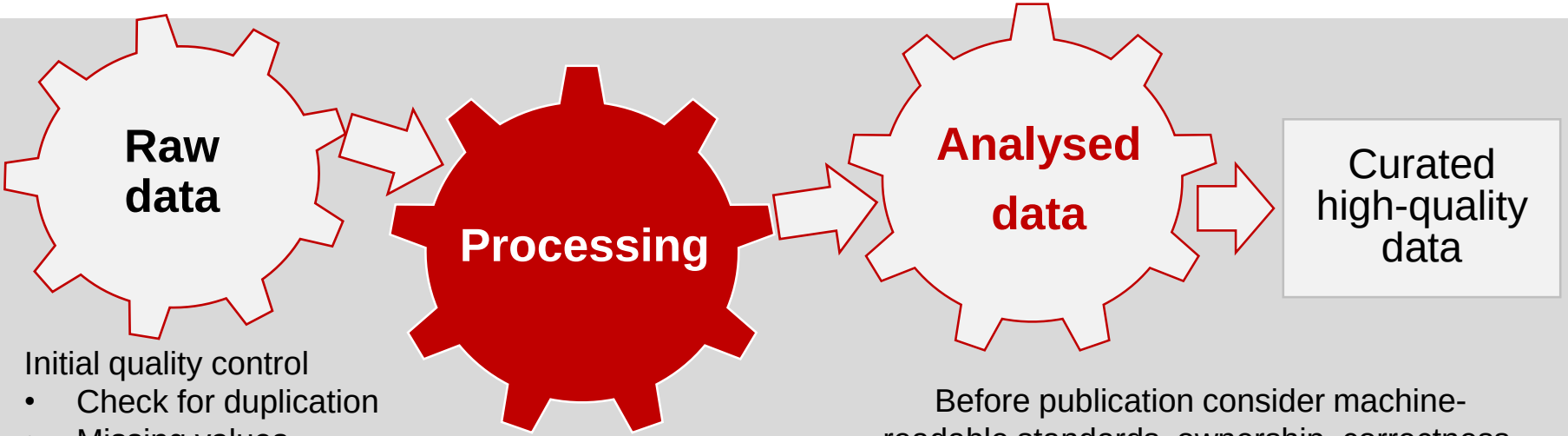
Legal requirements for IT security and data protection (ISDS analysis):

IT department (B. Hirschi)

Anonymization tools

https://arx.deidentifier.org/anonymization-tool/
https://amnesia.openaire.eu/

Data qualitiy phases



Initial quality control
- Check for duplication
- Missing values
- Tabular data
- Data protection
- Accuracy

Check for completeness, correctness, legislations and ethics, anonymize data

Before publication consider machine-readable standards, ownership, correctness, visualisation, legislations and ethics, codes and scripts cleaning, complete documentation

Metadata standards

UNIVERSITÄT BERN

Standards

- Metadata standards – recommended
- Machine-readable data formats

International standards where possible
(e.g., ISO19115 for geospatial metadata)

Dublin Core Metadata
Darwin Core Metadata Standard
North American Profile (NAP) of the ISO 19115
Metadata Schema 4.4

Metadata answer the following questions:

- **Who** created the data?
- **What** is the content of the data?
- **Why** were the data developed?
- **Where** is it geographically?
- **When** were the data created?
- **How** were the data developed?

→ Metadata are information that is needed to find and cite your data

UNIVERSITÄT BERN

Data documentation

Document, Discover and Interoperate (DDI)

- Standard for describing surveys, questionnaires, statistical data files, and social sciences information
- Used in: social, behavioral, economic, and health sciences
- Implemented in various repositories (e.g. FORS, Harvard Dataverse)
- DDI-Codebook and DDI-Lifecycle

Image: © 2018 DDI Alliance

Data documentation

Documentation = information that is needed to understand and re-use data

README file – [Example Paleoclimatology data]

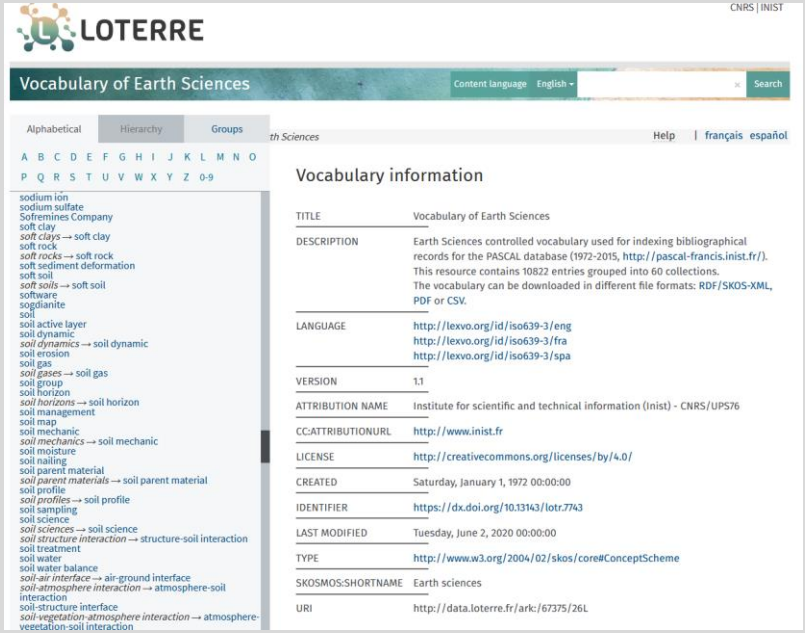
```
# Archive: Ice Core
#
# Dataset DOI:
#
# Parameter_Keywords: atmospheric gas, instrumental data
#--------------------
# Contribution_Date
#       Date: 2020-01-02
#--------------------
# File_Last_Modified_Date
#       Date: 2020-01-02
#--------------------
# Title
#       Study_Name: Polar Ice Core 42-77ka Excess Methane Data
#--------------------
# Investigators
#       Investigators: Lee, J.E.; Edwards, J.S.; Schmitt, J.; Fischer, H.; Bock, M.;
Brook, E.J.
#--------------------
# Description_Notes_and_Keywords
#       Description: All GISP2 CH4 measurements presented in Lee et al. (2020, GCA).
Includes previously published as well as unpublished values. Additional citations for
data include:
#       - Brook et al. (2000) "On the origin and timing of rapid changes in atmospheric
methane during the last glacial period." Global Biogeochemical Cycles, 14 (2), 559-572.
#               -The original data were on the old NOAA CMDL concentration scale.
#               -Corrected concentrations are given following Dlucockency et al. (2005)
to put the concentrations on the NOAA04 scale.
#               -These measurements were made at URI and WSU
#               -No solubility correction
```

Controlled vocabularies

Basic register of thesauri, ontologies & classifications
loterre.fr: Registry of controlled vocabularies

JISC: Directory of Metadata Vocabularies

Examples:
Medical Subject Headings (MeSH)
Astronomy Thesaurus
Thesaurus for Economics
Geographic Names® Online The Getty Research Institute
Linked **o**pen **ter**minology **re**sources



Codebook

A codebook communicates your research data to others, and ensures that the data can be properly understood and interpreted.

- Describes contents, structure, and layout of data collections
- Often used for tabular / statistical data

Show rows with cells including:

Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender

Image adapted from: Open Science Framework: How to make a data dictonary

Dataset versioning

Versioning

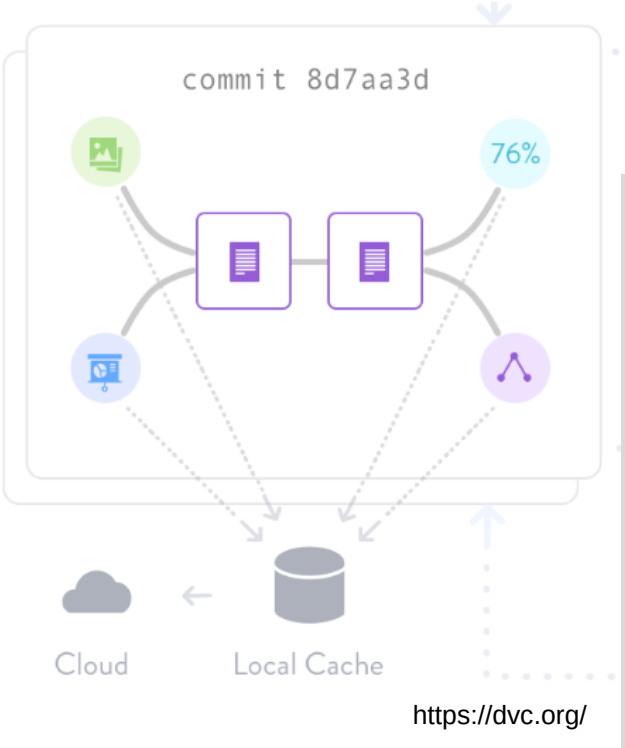
- Ticks with every publication describing the dataset
- Ticks every time the data change
- Ticks every time the metadata changed

Data Version Control (DVC)

[DVC does not replace Git!](#) DVC can be used as a Python library.

Git-compatible

Version control machine learning models, data sets and intermediate files.



https://dvc.org/

[Open-source Version Control System for Machine Learning Projects](#)

Changelogging

Backlog changes

Keeps track on changes to data and metadata, and why they were changed

- Changelogs
- Summaries for changes to compilation versions

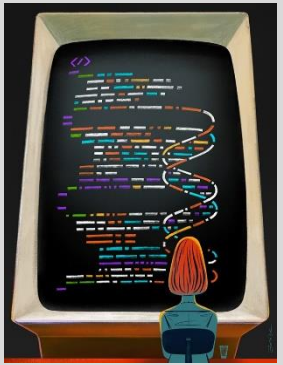


Illustration by Paweł Jońca

Backlog +…

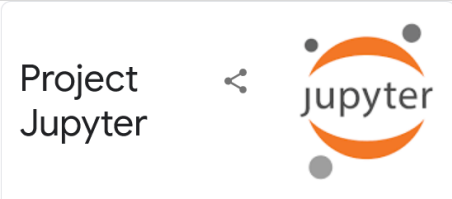
Classification +…

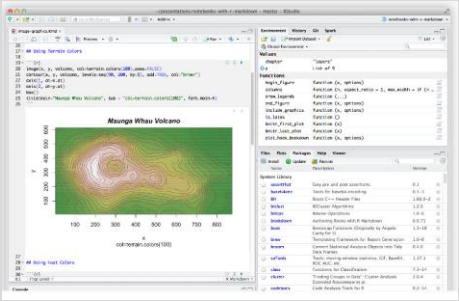
Ready for development +…

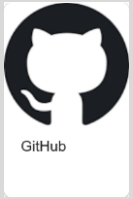
In progress..

https://github.com/

Computer code







- Comments in scripts
- Jupyter Notebook
- Data cleaning and transformation
- Statistical modeling
- Data visualization

Data cleaning

Data issues

- Missing values [0, 999] → NA
- Empty spaces
- URL, DOI links
- Remove irrelevant, duplicate and irregular values
- Data type
- Variation in units of measure
- Syntax (commas, length)
- Inconsistent values, typos

Correcting data

- Check, and adjust if not suitable
- Correct with regular expressions (regex)
- Cleaning Data in R - Video

Proxy_Metadata	If a data set has NOT been previously archived, enter "this study" in Original_Source_Data_URL. A template for that study will be generated.											
Missing_Value	NA											
dataID	FullDataSetCode	archiveType	latitude	longitude	elevation	startYear	endYear	Citation	Original_Source_D	variable_used	dat	
ID1	[ID1]_Mayotte_[Col1]_Mayotte_d18O	coral	-12.65	45.1	0	1865	1993	10.1007/s	https://www.ncdc	delta 18O		
ID11	[ID11]_Malindi_Kenya_[Col1]_Malindi_d18O	coral	-3	40	0	1801	1994	10.1126/s	https://www.ncdc	delta 18O		
ID16	[ID16]_Linsley_7_[Col1]_Rarotonga_d18O	coral	-21.2333	-159.833	0	1726	1996	10.1126/s	https://www.ncdc	delta 18O		
ID17	[ID17]_Linsley_6_[Col1]_Rarotonga_SrCa	coral	-21.2333	-159.833	0	1726	1996	10.1126/s	https://www.ncdc	Sr/Ca		
ID84	[ID84]_Taulis_1_[Col1]_Santiago de Chile_1	documentary	-33.3833	-70.7833	0	1540	1993	NA	this study	temperature	01.07.2017	Used in Best: 0
ID131	[ID131]_Plomo_Lake_[Col1]_Plomo_precip	paleolimnolog	-46.9833	-72.8667	203	1530	2000	10.1177/0	https://www.ncdc	precipitation	01.07.2017	Used in Best: 0
ID132	[ID132]_Pumacocha_d180_[Col1]_Pumacocha	paleolimnolog	-10.7	-76.0667	4300	-277	2007	10.1073/p	https://www.ncdc	delta 18O	01.07.2017	Used in Best: 0
ID140	[ID140]_Cascayunga_Peru_[Col1]_Cascayunga	speleothem	-6.06667	-77.2	885	1089	2005	10.1029/2	https://www.ncdc	delta 18O	01.07.2017	Used in Best: 0
ID141	[ID141]_Rasbury_1_[Col1]_Avaiki_ASM1	speleothem	-19	-169.833	NA	1829	2001	10.1029/2	https://www.ncdc	Couplet thickness	01.07.2017	Used in Best: 0
ID143	[ID143]_WA_Callitris_[Col6]_SW_Callitris_LKG	tree-ring	-33.0333	120.7667	300	1601	2005	NA	this study	trsgi	01.07.2017	Used in Best: 0
P4947	[IDP4947]_Ocean2kHR-PacificVanuatuKilbourn	coral	-15.7	167.2	NA	1928	1992	10.1029/2	https://www.ncdc	delta 18O	01.07.2017	Used in Best: 0
P4948	[IDP4948]_Ocean2kHR-PacificVanuatuKilbourn	coral	-15.7	167.2	NA	1928	1992	10.1029/2	https://www.ncdc	Sr/Ca	01.07.2017	Used in Best: 0
P4983	[IDP4983]_SAm-QuelccayaIce.Thompson.2013_	ice core	-13.9333	-70.8333	NA	226	2009	10.1126/s	https://www.ncdc	delta 18O	01.07.2017	Used in Best: 0
P4984	[IDP4984]_SAm-QuelccayaIceThompson2013_	ice core	-13.9333	-70.8333	NA	226	2009	10.1126/s	https://www.ncdc	ice accumulation r	01.07.2017	Used in Best: 0
SAm_3	Laguna Aculeo_PAGES2013	paleolimnolog	-33.14	-70.15	NA	NA	NA	10.1007/s	https://www.ncdc	pigment reflection	01.07.2017	Used in Best: 0
SAm_4	Cariaco Basin_PAGES2013	paleoceanogra	9.88	-64.13	NA	NA	NA	10.1029/2	https://www.ncdc	Mg/Ca	01.07.2017	Used in Best: 0
SAm_23	Rio Gallegos DJF_PAGES2013	Instrumental	-52	-69.45	NA	NA	NA	10.1007/s	https://www.ncdc	temperature	01.07.2017	Used in Best: 0

Konecky et al. 2020 ESSD

Good Laboratory Practice (GLP)

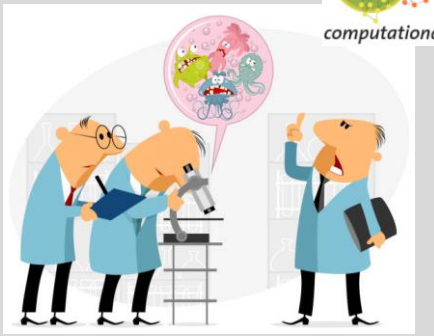
Non-clinical studies

- Pharmaceuticals
- Animals
- Environment

Data quality control

- Laboratory protocols, Electronic Laboratory Notebooks (e.g. RSpace ELNs), OpenBIS
- Versioning
- Codebook

csb computational systems biology



Good Clinical Practice (GCP)

Data quality control

- Protocol documentation
- Data protection and legislation
- Ethics committees
- Anonymisation/Pseudonymisation

Good Clinical Practice:
Integrated Addendum to ICH E6(R1) FDA E6 GCP

Kanton Bern
Canton de Berne

swissmedic

UNIVERSITÄT BERN

Data quality and metadata standards

References and useful links

Cleaning Data in R – video tutorial

Introduction to ELNs and LIMS, DLCM Guidelines

openBIS User Group Meeting 2021

Good laboratory practice compliance

OECD: Good laboratory practice
European Commission: Good laboratory practice

EU GLP Working Group

Princeton University Data and Statistical Services, “How to Use a Codebook”

Data quality and metadata standards

Support

- Data management plan review

Submit online or via E-Mail

- Research data management trainings and courses

Website link

Open Science News

Data Management Plan Review

Open Science

UNIVERSITÄT BERN

Data quality and metadata standards

Support

BORIS Portal Research Data, Projects and Fundings

BORIS Portal Training and Workshops (Link)

Prof. Dr. Christian Leumann, Rector of the University of Bern, on BORIS Portal, research data and project data (Link).

Borisportal@ub.unibe.ch

Thank you

for your attention!

Open Science Team
E-Mail: openscience@ub.unibe.ch